Автоматизированная разметка набора открытых данных с применением больших языковых моделей

Комольцев Данил Алексеевич, студент магистратуры
Научный руководитель: Васендина Ирина Сергеевна, кандидат технических наук, доцент
Северный (Арктический) федеральный университет имени М. В. Ломоносова (г. Архангельск)

В статье автор описывает процесс автоматизированной разметки набора текстовых данных посвящённых тематике вакцинации с применением больших языковых моделей.

Ключевые слова: *анализ больших данных, компьютерная лингвистика, большие языковые модели, llm-модели.*

В рамках работы с большими текстовыми данными частой проблемой является получение размеченных данных. Первичные данные могут содержать тексты, тематика которых не соответствует требованиям исследования. Существует несколько способов разметки данных. Таким способом может быть ручная разметка. Эксперт или несколько экспертов вручную размечают набор текстовых данных, принимая решение об их релевантности. Этот способ, пускай и может считаться надёжным, обладает рядом ограничений, которые осложняют его применение для большого объёма данных.

Альтернативой ручной разметке может быть автоматизированная разметка. В настоящее время большие языковые модели (LLM-модели) могут упростить и ускорить разметку текстов. Разметка текста является задачей классификации, которую LLM-модели решают весьма успешно [1] [3].

В рамках магистерского исследования была проведена автоматизированная разметка датасета, собранного на основе открытых данных социальной сети «Вконтакте». Часть набора данных представлена на рисунке 1.

Набор данных собирался с помощью запросов по ключевым словам с помощью API Вконтакте. В наборе данных присутствуют как тексты, посвящённые тематике вакцинации, так и тексты, которые тематике не соответствуют.

Для верификации тематики использовалась llm-модель mixtral-8x7b, размещённая на ресурсах сервиса Groq Cloud. В архитектуре mistral, на которой основывается вышеупомянутая llm-модель, используются технологии сосредоточенного внимания по группам запросов (Grouped-query attention, GQA) и скользящего окна внимания (Sliding window attention, SWA), что позволяет обрабатывать большие запросы с меньшими затратами на вычисление [2].

Для того, чтобы обеспечить работоспособность модели был создан технический промпт — набор инструкций, который определяет ожидаемый результат работы модели. Технический промпт представлен на рисунке 2.

Данный промпт указывает модели, что тексты необходимо сортировать по 6 категориям в зависимости от содержания текстов, а также что модели необходимо объяснить, почему была выбрана именно это категория. Результат необходимо оформить как json.

Запросы к Groq Cloud отправляются с помощью библиотеки groq и библиотеки requests для языка программирования python. Листинг запроса представлен на рисунке 3.

При работе с API существует вероятность ошибок. Именно поэтому был написан алгоритм обработки ошибок, представленный на рисунке 4.

Алгоритм верификации построен следующим образом. На вход модели подаётся pandas-датафрейм, содержащий не более 14400 строк, итеративно обрабатываются строки: извлекается текстовая информация, которая вместе с системным пром-

	text	id	date	link
0	По традиции делимся полезной информацией, кото...	558185313	2024-04-10 05:00:14	https://vk.com/wall558185313_769
1	👍Преимущества ЧАЯ ДЛЯ ПЕЧЕНИ👉»ФанДетокса»:\n\n...	783459251	2024-04-10 04:58:58	https://vk.com/wall783459251_364
2	10.09.2024 г. в ДНП провели лекцию на тему бол...	-216942256	2024-04-10 04:57:28	https://vk.com/wall-216942256_131
3	Профилактика острых кишечных инфекций \n В пер...	635313754	2024-04-10 04:50:19	https://vk.com/wall635313754_1160
4	ЧАРЛАЛ!!!\nЧаа-Хол коңуунун топ эмнелгезинин у...	-104541079	2024-04-10 04:49:02	https://vk.com/wall-104541079_41380
...	...	...	...	...
72185	ТГ ЗВЕЗДЫ\n\n1. ❗ Уголовное расследование в от...	-95865483	2024-04-04 13:15:00	https://vk.com/wall-95865483_432481
72186	Европейская прокуратура открыла расследование ...	-34777837	2024-04-04 12:16:00	https://vk.com/wall-34777837_2685122
72187	Евросоюз: коррупция и смерти в "райском саду"\...	-167661899	2024-04-03 23:20:00	https://vk.com/wall-167661899_2007481
72188	В период так называемой пандемии я активно выс...	355949337	2024-04-03 22:46:57	https://vk.com/wall355949337_32862
72189	В Жогорку Кенеше (парламенте Кыргызстана) 3 ап...	-89154218	2024-04-03 13:37:22	https://vk.com/wall-89154218_47489

72190 rows × 4 columns

Рис. 1. **Набор данных из социальной сети «Вконтакте»**

```
[ ] system_prompt = """Твоя задача проанализировать текст и предположить, посвящён ли данный текст тематике вакцинации людей или животных в прямом и только в прямом смысле.
    Тебе необходимо дать ответ в формате json. Не нужно добавлять никаких дополнительных символов вроде "\n"

    {"estimate": 1,
    "proof": Предложение, обосновывающее выбор}

    В поле 'estimate' тебе необходимо предоставить значение

    0 - речь не о вакцинации
    1 - речь исключительно о вакцинации людей
    2 - речь о вакцинации животных
    3 - речь о конспирологических теориях, касающихся тематики вакцинации
    4 - В тексте упоминаются заболевания, но нет упоминания вакцинации как таковой
    -1 - текст не предоставлен, не был распознан, произошла ошибка

    В поле 'proof' тебе необходимо предоставить обоснование для значения 'estimate'
    """
```

Рис. 2. **Технический промпт для llm-модели mixtral-8x7b**

Groq

```
# @title Groq
model_groq = "mixtral-8x7b-32768" # @param ["mixtral-8x7b-32768", "gemma-7b-it", "llama3-70b-8192", "llama3-8b
groq_api_key = "-" # @param {type:"string"}
def groq_analyse(text, groq_api_key):


  time_out = 2 # @param {type:"slider", min:0, max:10, step:1}
  client = Groq(api_key=groq_api_key, timeout=time_out)
  completion = client.chat.completions.create(
    model=f"{model_groq}",
    messages=[
        {
            "role": "system",
            "content": f"{system_prompt}"
        },
        {
            "role": "user",
            "content": f"{text}"
        }
    ],
    temperature=0,
    max_tokens=30000,
    top_p=1,
    stream=False,
    response_format={"type": "json_object"},
    stop=None,
  )
  return completion.choices[0].message.content
```

Рис. 3. **Запрос к Groq Cloud**

Обработка ошибок

```
[ ] # @title Обработка ошибок

    def analyse_with_backup(text):
      wait_time = 4 # @param {type:"slider", min:0, max:10, step:1}
      try:
        print(f"Используется модель по-умолчанию ({model_groq})")
        result = groq_analyse(text, groq_api_key)
      except groq.APITimeoutError:
        print(f"API не отвечает. Попытка снова через {wait_time}")
        time.sleep(wait_time)
        result = analyse_with_backup(text)
      except groq.BadRequestError:
        print(f"Используется запасная модель ({model_backup})")
        result = groq_analyse_backup(text, groq_api_key)
      except groq.RateLimitError:
        print(f"Превышен лимит запросов. Попытка снова через {wait_time}")
        time.sleep(wait_time)
        result = analyse_with_backup(text)
      except groq.InternalServerError:
        print(f"Сервер лежит, попытка снова через {wait_time} секунды")
        time.sleep(wait_time)
        result = analyse_with_backup(text)
      return result
```

Рис. 4. **Алгоритм обработки ошибок**

птом формирует запрос к llm-модели. На выходе получается либо ошибка, в таком случае запрос отправляется снова спустя какое-то время, либо ответ в формате json, который в свою очередь добавляется в результирующий датафрейм. Датафрейм сохраняется на Google Drive автоматически. Также реализована возможность продолжить обработку с итерации, на которой обработка прервалась даже спустя продолжительное время. Листинг алгоритма верификации представлен на рисунке 5.

Обработка набора данных заняла около 504 часов. Результирующий датасет представлен на рисунке 6.

Основным фактором, влияющим на скорость обработки, были ограничения платформы google colab, на которых запускался алгоритм и ошибками на стороне серверов Groq Cloud.

```python
# @title Анализ
top = 0 # @param {type:"integer"}
bottom = top+14400
start_from_broken_iteration = True # @param {type:"boolean"}

if start_from_broken_iteration == True:
  top = top+len(df_result)
df_cut = df[top:bottom]
df_cut['text'].fillna('no text', inplace=True)
df_cut


path_to_save = "/content/gdrive/MyDrive/Posts_evaluation/" # @param {type:"string"}
save_after_iteration = 10 # @param {type:"integer"}

iter = 0
max_allowed_tokens = 4000
if start_from_broken_iteration == True:
  iter = len(df_result)



for index, row in df_cut.iterrows():
    text = row['text']
    link = row['link']
    if iter == 0:
        safe_text = truncate_to_token_limit(text, max_allowed_tokens)
        print(f"===\n{safe_text}\n===\n")
        analysis = analyse_with_backup(safe_text)
        analysis_json = try_parse_json(analysis)
        analysis_json.update({'text': text, 'link': link})
        df_result = pd.DataFrame([analysis_json])
    else:
        safe_text = truncate_to_token_limit(text, max_allowed_tokens)
        print(f"===\n{safe_text}\n===\n")
        analysis = analyse_with_backup(safe_text)
        new_row = try_parse_json(analysis)
        new_row.update({'text': text, 'link': link})
        new_row_df = pd.DataFrame([new_row])
        df_result = pd.concat([df_result, new_row_df], ignore_index=True)
    iter += 1
    if iter % save_after_iteration == 10:
        print("Result Autosaved")
        df_result.to_csv(path_to_save+f"autosave_14400.csv", index=False)
```

Рис. 5. **Алгоритм выставления метки**

	estimate	proof	text	link
0	0	Текст не содержит упоминаний вакцинации людей ...	По традиции делимся полезной информацией, кото...	https://vk.com/wall558185313_769
1	0	Текст не содержит упоминаний вакцинации людей ...	👍 Преимущества ЧАЯ ДЛЯ ПЕЧЕНИ👉»ФанДетокса»:\n\n...	https://vk.com/wall783459251_364
2	4	The text discusses the topic of hepatitis A, o...	10.09.2024 г. в ДНП провели лекцию на тему бол...	https://vk.com/wall-216942256_131
3	4	The text discusses diseases such as acute inte...	Профилактика острых кишечных инфекций \n В пер...	https://vk.com/wall635313754_1160
4	0	The text does not mention vaccination or disea...	ЧАРЛАЛІІІ\nЧаа-Хол коңуунун топ эмнелгезинин у...	https://vk.com/wall-104541079_41380
...	...	...	...	...
72185	1	The text is about the investigation of Ursula ...	ТГ ЗВЕЗДЫ\n\n1. ❗ Уголовное расследование в от...	https://vk.com/wall-95865483_432481
72186	3	The text discusses allegations of corruption a...	Европейская прокуратура открыла расследование ...	https://vk.com/wall-34777837_2685122
72187	0	Текст не содержит упоминаний вакцинации людей ...	Евросоюз: коррупция и смерти в "райском саду"\...	https://vk.com/wall-167661899_2007481
72188	3	The text discusses a conspiracy theory related...	В период так называемой пандемии я активно выс...	https://vk.com/wall355949337_32862
72189	0	The text does not mention vaccination or any r...	В Жогорку Кенеше (парламенте Кыргызстана) 3 ап...	https://vk.com/wall-89154218_47489

72190 rows × 4 columns

Рис. 6. **Список постов с меткой релевантности**